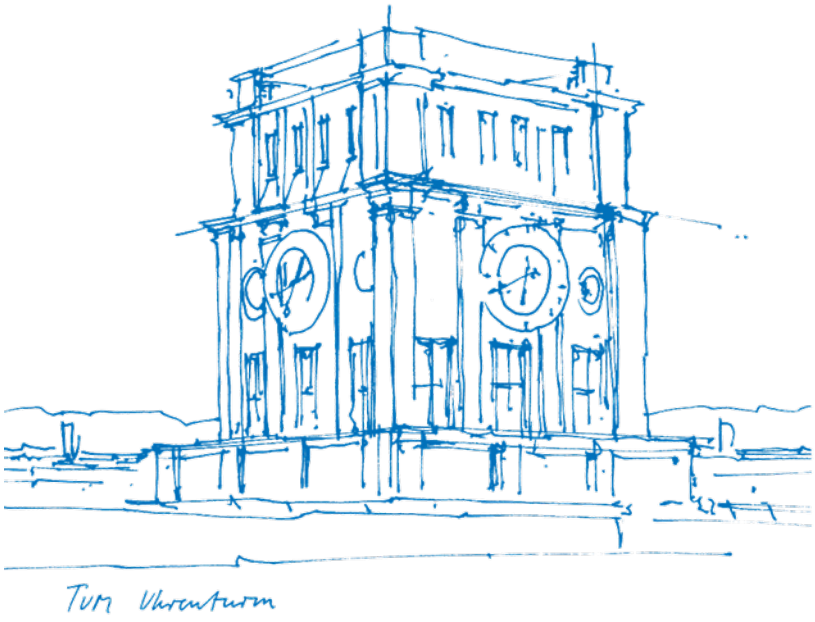
Applied Econometrics - Lecture 2

Prof. Dr. Helmut Farbmacher
Technical University of Munich
TUM School of Management
Munich, April 8, 2025



Linear Regression Model

Outline of today's lecture

Linear Regression Model

- Assumptions of OLS
- Example: Private Returns to Education
- Identification and Estimation of Parameters
- Ceteris Paribus
- Asymptotic Properties of OLS

The Population Model

$$y_i = \beta x_i + u_i$$

Example: Private returns to education; y (income) and x (years of schooling)

We want to use **data** and **credible assumptions** to estimate β.

The most important assumption is a statement about u, which we need for consistency of our estimate.
Recall: Consistency means that the estimate falls close to the true value (if n is large).

↳ $\hat{\beta} \to \beta$ vs. unbiasedness

These assumptions are at the heart of econometrics. Defending them is an important task in any empirical study.

Linear Regression Model

Assumptions of OLS

Ordinary Linear System

Ordinary least squares

Assumptions of OLS

$$y_i = \beta x_i + u_i$$

Assumptions of OLS:

OLS.1: $\mathrm{E}(x_i u_i) = 0$

We do not want a covariance between x and u.
this equation ensures that

OLS.2: $\mathrm{E}(x_i^2) \neq 0$

Variance of x when mean is 0
-> if variance is zero x is always the same

Assumptions of OLS

OLS.1: $\mathrm{E}(x_i u_i) = 0$

Remember the definition of covariances: $\mathrm{Cov}(x_i, u_i) = \mathrm{E}(x_i u_i) - \mathrm{E}(x_i)\mathrm{E}(u_i)$

OLS.1 is a statement about the covariance between x and u since wlog $\mathrm{E}(u_i) = 0$.
If $E(u_i)$ is indeed unequal to zero in the population, it would end up in the constant.

OLS.1 is **untestable** because we do not observe u in our data.
It is the **identifying assumption** of OLS estimation.

If $\mathrm{E}(x_i u_i) = 0$, we call x_i an **exogenous** regressor. good
If $\mathrm{E}(x_i u_i) \neq 0$, we call x_i an **endogenous** regressor. bad

Assumptions of OLS

To defend OLS.1, we need to think about

1. which variables could be in the error term u **and**
2. whether these variables are correlated with x

We call a variable an **omitted variable** if this variable

1. is in the error term u (i.e. has an effect on the outcome y) **and**
2. is correlated with x

e.g. height and income?

OLS.1 is violated (i.e., x is endogenous) if an omitted variable exists.

Omitted variables are a huge problem in empirical economics (particularly, when we analyze observational data).

Assumptions of OLS

OLS.2: $\mathrm{E}(x_i^2) \neq 0$

Remember the definition of variances: $\mathrm{Var}(x_i) = \mathrm{E}(x_i^2) - \mathrm{E}(x_i)^2$

OLS.2 is a statement about the variance of x since wlog $\mathrm{E}(x_i) = 0$.
If $\mathrm{E}(x_i)$ is indeed unequal to zero in the population (say, $= 2$), we could think about a new variable $\widetilde{x}_i = x_i - 2$ and get $\mathrm{E}(\widetilde{x}_i) = \mathrm{E}(x_i) - 2 = 0$.

OLS.2 depends only on observed variables (i.e. it is verifiable) and just rules out that x_i is a constant.

OLS.2 is a technical assumption (without any economic content). Apart from the obvious fact that x_i cannot explain changes in y_i if x_i itself does *not* vary.

Assumptions of OLS

$$y_i = \beta x_i + u_i$$

If OLS.1 and OLS.2 are fulfilled, we can interpret β as the **causal effect** of x on y.

Causal effect: The expected change of the outcome y when we randomly draw an object (e.g. a person, firm etc) from the population and change x by one unit.

For example,

- increasing price by one unit & its causal effect on sales
- increasing compulsory schooling by one year & its causal effect on earnings (private returns to education) or crime (social returns to education).

Assumptions of OLS

OLS assumptions can be generalized to more than one regressor with (essentially) the same interpretation.

Notation, however, needs more sophisticated matrix algebra, which I am not going to introduce in this course.

If you want to know more, join my lecture “Microeconometric Methods for Big Data” (Usually every winter term. It’s a Master class but can also be attended for credits at Bachelor’s level).

Linear Regression Model

Example: Private Returns to Education

Example: Private Returns to Education

$$y_i = \beta x_i + u_i$$

Private returns to education: y **income** and x **years of schooling**

Interesting research question: How much more can you expect to earn with one additional year of schooling?

Of course, there may be other factors that influence the outcome of y (e.g., income also depends on job experience or soft skills).

For now, the above linear regression assumes that these other factors are collected in the error term, u.

Omitted variables

Example: Private Returns to Education

$$\text{inc}_i = \beta \text{educ}_i + u_i$$

OLS.1: $\mathrm{E}(\text{educ}_i\, u_i) = 0$

Can we think about a variable in u that is correlated with *educ*?

Potential **omitted variable** (for example, *social skills*):

1. Employers are willing to pay on average higher wages to people with better social skills (i.e. social skills is in u_i) **and**
2. People with better social skills might also be more successful in school on average (i.e. social skills is correlated with education).

We will discuss omitted variables in more detail in lecture 3.

Example: Private Returns to Education

$$\text{inc}_i = \beta \text{educ}_i + u_i$$

OLS.2: $\mathrm{E}(\text{educ}_i^2) \neq 0$

Can be easily verified (since it only depends on an observed variable, *educ*)

Just check whether there is variation in *educ*.

For illustration here some examples of a **violation**:

- We only have observations with 12 years of education in our sample.
- Or we want to estimate the gender wage gap but only have women in our sample.

Linear Regression Model

Identification and Estimation of Parameters

Identification and Estimation of Parameters

$$\begin{aligned} y_i &= \beta x_i + u_i \\ \Leftrightarrow \quad u_i &= y_i - \beta x_i \end{aligned}$$

for linear ops you can split, otherwise only if independent

Expection Value

If OLS.1 and OLS.2 hold, we can identify β

$$\begin{aligned} \mathrm{E}(x_i u_i) &= E\left(x_i (y_i - \beta x_i)\right) \\ &= \mathrm{E}\left(x_i y_i - \beta x_i^2\right) \\ &= \mathrm{E}(x_i y_i) - \beta \mathrm{E}(x_i^2) = 0 \end{aligned}$$

OLS.1

rearranging terms

$$\beta = \frac{\mathrm{E}(x_i y_i)}{\mathrm{E}(x_i^2)}$$

Now, you also see why we need OLS.2: $\mathrm{E}(x_i^2) \neq 0$.

Identification and Estimation of Parameters

If OLS.1 and OLS.2 hold, we can identify β

$$\beta = \frac{\mathrm{E}(x_i y_i)}{\mathrm{E}(x_i^2)}$$

β is for population,

$\hat{\beta}$ is for the sample

sample mean $\Leftrightarrow$ E

Estimation is a piece of cake (just replace expectations by sample means).

Let $\widehat{\beta}$ denote the OLS estimate of β,

$$\widehat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$\frac{X \cdot Y}{X \cdot X}$ why is this beta?

Of course, $\frac{1}{n}$ is a constant in any particular dataset and cancels out.

$\widehat{\beta}$ is a (clever) function of the observed variables y and x.

Linear Regression Model

Ceteris Paribus

Ceteris Paribus

So far, we considered linear regressions with only one explanatory variable:

$$wage_i = \beta_0 + \beta_1 educ_i + \varepsilon_i$$

Say, people with more education can expect to earn higher wages (i.e. $\beta_1 > 0$).

Confounding effects: For example, ...
People with higher education also have a higher level of social skills.
Employers may also be willing to pay higher wages for higher level of social skills.

Solution: We can control for additional variables (like social skills). That is we include those variables in our regression.

Big advantage: Every variable that we control for in our regression is not in the error term *u* and, hence, **cannot** cause an omitted variable bias.

Ceteris Paribus

OLS 1 = educ; ε_i

$$wage_i = \beta_0 + \beta_1 educ_i + \underbrace{\beta_2 skills_i + u_i}_{\varepsilon_i}$$

ε_i ← error term

Controlling for additional variables allows a **ceteris paribus** interpretation of the estimates.

For example, **all else equal** the private returns to education is 5000€ for every additional year of schooling.

We often also say: “we hold these other factors/variables fixed or constant”.

Note that “all else equal” is only referring to variables in our regression.

Ceteris Paribus

Suppose we run two regressions on the same outcome variable y:

(a bivariate regression) $$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$$ ← error-term

(a multivariate regression) $$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + u_i$$

The error term ε_i in the bivariate regression is then a composite error term that contains observed and unobserved variables:

$$\varepsilon_i = \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + u_i$$

For OLS.1 to be fulfilled in the **bivariate** regression, x_1 needs to be uncorrelated with the *observed* variables $(x_2, \ldots, x_p)$ and the *unobserved* variables in u.

For OLS.1 to be fulfilled in the **multivariate** regression, x_1 being uncorrelated with the *unobserved* variables in u is sufficient (of course, the same should hold for $x_2, \ldots, x_p$).

Ceteris Paribus

OLS.2 in the multivariate regression is again a technical assumption similar as in the linear bivariate regression.

It still rules out that any x_j is a constant.

Additionally, the assumption now rules out perfect multicollinearity.
Perfect multicollinearity refers to the fact that an x_j can be perfectly predicted by a (deterministic) linear combination of the other regressors in our model.

Example:

$$y = \beta_0 + \beta_1 \text{male} + \beta_2 \text{female} + u$$

we can predict male from female

Ceteris Paribus: An Example

Data info:

The Current Population Survey (*cps09mar*) contains information on employment, earnings, educational attainment, income etc. for 57.000 U.S. households (March 2009).

Variables:

earnings	total annual earnings
education	years of education (based on highest degree)
hours	number of hours worked per week

Data can be downloaded here: `https://www.ssc.wisc.edu/~bhansen/econometrics/`.

Ceteris Paribus: An Example

$$earnings_i = \beta_0 + \beta_1 educ_i + \varepsilon_i$$

$$lm(formula = earnings \sim education, data = cps09mar)$$

```
MODEL INFO:
Observations: 50742
Dependent Variable: earnings
Type: OLS linear regression

MODEL FIT:
F(1,50740) = 8796.52, p = 0.00
R² = 0.15
Adj. R² = 0.15

Standard errors: OLS
------------------------------------------------------------
                         Est.      S.E.   t val.      p
----------------- ----------- --------- -------- ------
(Intercept)         -46755.01   1106.79   -42.24   0.00
education            7314.13     77.98    93.79   0.00
------------------------------------------------------------
```

$\widehat{\beta}_1 = 7314.13$

Ceteris Paribus: An Example

$$earnings_i = \beta_0 + \beta_1 educ_i + \beta_2 hours_i + u_i$$

$$lm(formula = earnings \sim education + hours, data = cps09mar)$$

```
MODEL INFO:
Observations: 50742
Dependent Variable: earnings
Type: OLS linear regression

MODEL FIT:
F(2,50739) = 6019.91, p = 0.00
R² = 0.19
Adj. R² = 0.19

Standard errors: OLS
-------------------------------------------------------
                         Est.      S.E.   t val.      p
---------------- ------------ --------- -------- ------
(Intercept)        -102039.32   1505.79   -67.76   0.00
education             6765.70     76.66    88.26   0.00
hours                 1435.66     27.31    52.58   0.00
-------------------------------------------------------
```

$\widehat{\beta}_1 = 6765.70$

Ceteris Paribus: An Example

Interpretation of regression output:

We observe that an additional year of education goes along with $\widehat{\beta}_1 = 6765.70€$ higher earnings on average (controlling for hours worked).

Is this result significant different from zero? Test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ using a t-test:

$$t = \frac{\widehat{\beta}_1 - 0}{\mathrm{SE}_{\widehat{\beta}_1}} = 88.26$$

Since $|t| \gg 2.57$ (1% level), we conclude that it is significant at any conventional level.

However, is this a causal result?

Linear Regression Model

Asymptotic Properties of OLS

Asymptotic Properties of OLS

If the sample size becomes large (technically, $n \to \infty$) and OLS.1 and OLS.2 hold, we can show the following properties of OLS:

1. **Consistency:** $\widehat{\beta}$ converges in probability to the true values β, we write

$$\widehat{\beta} \xrightarrow{p} \beta$$

2. **Asymptotic normality:** The distribution of $\widehat{\beta}$ converges to a normal distribution, we write

$$\widehat{\beta} \xrightarrow{d} N\left(\beta, \text{Var}\left(\widehat{\beta}\right)\right)$$

The proofs of both are beyond this course. But you should know the implications of both results. We always use them in applied research.

Asymptotic Properties of OLS

Consistency is an important property of estimators. We want OLS to be consistent, i.e. we want our estimates to get closer and closer to the true values if the sample size increases. This is why we have to check and defend the OLS assumptions.

Consistency, however, does not tell us anything about how $\widehat{\beta}$ behaves in finite samples.

Here **asymptotic normality** and the **variance** of $\widehat{\beta}$ comes into play. It tells us in what range we expect β to be, which allows us to perform hypothesis tests.

Asymptotic results are valid for large samples only. How large?
Hard to say in general. Ongoing research! If you are interested, see the discussion about randomization inference in chapter 4.2 (not relevant for the exam):
`https://mixtape.scunning.com/index.html`

Asymptotic Properties of OLS

Under homoskedasticity, the **variance** of $\widehat{\beta}$ in the **bivariate** regression model is given by

$$\mathrm{Var}(\widehat{\beta}) = \frac{1}{n}\frac{\widehat{\sigma}_u^2}{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \qquad \text{and} \qquad \mathrm{SE}(\widehat{\beta}) = \sqrt{\mathrm{Var}(\widehat{\beta})}$$

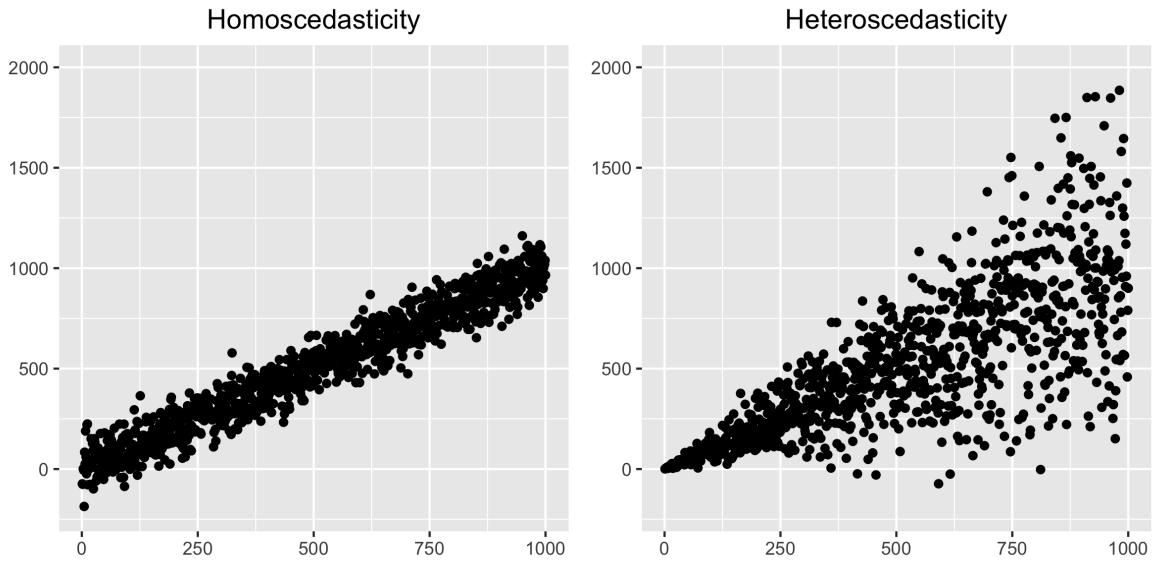
with $\widehat{\sigma}_u^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i^2$ and $\widehat{u}_i = y_i - \widehat{\beta} x_i$.

That is we can more **precisely** estimate $\widehat{\beta}$ if

- the model variance $\widehat{\sigma}_u^2$ decreases
- the sample sizes n or the variance of x increases.

The formula for the **multivariate** regression model needs more sophisticated matrix algebra. The takeaways are however very similar to the bivariate case.

Sidenote: Homoskedasticity vs Heteroskedasticity



Statement about whether the variance of u_i changes with the level of x_i.

Most economic outcomes are heteroskedastic. Heteroskedasticity-robust standard errors exist and are implemented in all statistical packages.

Recommended reading

For next week please read chapter:

8.2 Panel Data: Estimation

in `https://mixtape.scunning.com/index.html`

Contact

Helmut Farbmacher
office.econometrics@mgt.tum.de